5-1988

Inter-Rater Reliability: A Question of Measurement in Social Science Research

Lani Kai Eggertsen Goff
Utah State University

Inter-Rater Reliability:

A Question of Measurement in Social Science Research

A description of the work
completed to fulfill
the requirements
for Honor's Senior Project

by

Lani Kai Eggertsen Goff

Spring 1988

Erik Erikson (1968) set forth the framework for the conceptualization and adolescent identity formation in the late 1960's. The assessment of identity invokes a subjective measurement of a youth's process and current standing regarding sociopsychological development. According to Adams (1987) throughout Erikson's writings a "conscious sense of individual identity," a process of "ego synthesis" and formation of sense of social "ideals and social identity" are central considerations when discussing identity.

The measurement of identity has been operationalized by Marcia (1966) and Matteson (1977). Gerald Adams (1987) states that assessment includes "two dimensions involving the presence or absence of" an "exploration period" and a "clearly defined and stable commitment to values, beliefs, and standards." It is expected of youths that they experience a "crisis period" or "a psychosocial moratorium wherein the adolescent is expected to make 'commitments for life'," and "a relatively fixed self-definition" (Adams, 1987, p.3).

The two methods most frequently used to measure identity status have been self reporting scales and clinical interviews. Both involve categorization of youths into four statuses. Those who have not made a commitment for experienced "a compulsion to explore life alternatives" are termed identity diffused. Another category is identity foreclosed. The youth who has acquired commitments from others and has not tested their stated commitments are called "foreclosed." One who is currently

exploring but is not committed to any self-defined identity is termed as moratorium. The final category includes youths who have experienced a moratorium type stage and made "substantial exploration prior to identifying personal and unique ideological commitments" and is called identity achieved (Adams, 1987, p. 4).

In measuring the levels of commitment and exploration a clinical interview is coded by two individuals to obtain an inter-rater reliability (consensual validation) (Adams, 1987, p.5). This subjective agreement on how to evaluate the responses of the adolescents is an attempt to make a scale that will objectively measure identity formation.

The Utah Parent-Adolescent Project is an extensive study being conducted under the supervision of Dr. Gerald Adams. Many different instruments are being used to gather data from sixty families in Utah. The families were selected with the considerations of how likely it was that they would be willing participants for the three year study. Each family had to have a child between the ages of 14 and 16 in order to be included in the study. A battery of questionnaires were administered to the father, the mother, and the adolescent. In addition each family was visited by a trained team of interviewers to administer instruments to all three family members.

The families also were part of a 45 minute to one hour identity interview. The interview is designed to measure the individual's status on a two dimensional model illustrated in Figure 1 (Adams, Lee, and Bennion, 1987, p.5). The amount of

exploration and commitment to a particular domain is measured on a scale of one through four but not by fractions of whole numbers, i.e. only a 1, 2, 3, or 4 could be assigned to the interviewees level of exploration and commitment.

Each interview included eight domains, Occupation, Politics, Religion, Philosophical Lifestyles, Friendship, Dating (or Parenting), Recreation, and Gender Roles. The interviewer gave a brief introduction and/or definition of the domain and then depending on the responses had a structured set of questions to ask. The objective of the interviewer was to obtain sufficient information for a coder to be able to classify the individual into one of the statuses.

My role in the project was that of coder of all the adolescent interviews. The project, or the interview portion of the project, was explained to me and a training session scheduled. Another student was chosen to code also and the task of learning the scales, listening to practice tapes and then actually assigning a status to the individual for each domain began.

The purpose of having two coders was to establish an estimate of reliability. It became quite clear to me that the objective of pigeon-holing each adolescent into a status based on their responses wasn't as simple or clear cut as one might think. The measure of inter-rater reliability is used to determine whether the data can be evaluated by different coders and have a high percentage of agreement between them. It is usually

accepted that the data is significant if the inter-rater reliability is about 75 percent agreement. However many such studies have inter-rater reliability rating ranging from "48 percent to 78 percent" (Adams, 1988).

One of the questions this raised for me was "who set this standard and why did they choose that percentage?" I am concerned that research is ignored because the "statistics show" that the data is not reliable. The problems of insufficient data, poor training, failure of the recorders of even the personality style of the adolescent could be factors that effect the final conclusions of those who conduct research but are not recognized as flaws. Instead the study may be considered to be insignificant or not worth replicating because the inter-rater reliability is too low.

Any group can, with enough deliberation, come to a conclusion as to which factors of the interview determine what status the adolescent is in. A young man age thirteen is highly unlikely to have studied, much less even thought about, his beliefs on gender roles. However, he may have thought extensively about it but couldn't express adequately his thoughts. In this particular study there is no way of knowing whether the person is not able to express himself or whether the interviewer just didn't ask the questions in a way he could relate to and respond to.

The percentage of status rating agreement between our coding was 70.49 percent. That is, 29.51 percent of the ratings we gave

to each adolescent as independent raters were different. This is considered by most social science statistics to be an acceptable range of inter-rater reliability.

Problems we saw with this were that the interviewers errors were not apparent and not revealed by the inter-rater reliability measure. Examples of this include: often times they did not define the domains properly, did not ask enough questions for us to determine what the adolescents status was. In a few interviews domains were entirely skipped.

We tried to take the problems in interviewing into consideration by reviewing the interviews and identifying which of the status placement decisions were guesses. Of the approximately thirty percent inter-rater discrepancies 43.75 percent were guesses. This means that close to fifty percent of the items on which we as coders disagreed did not have sufficient information, even no information at times, for us to base our decisions on.

Finally, in order to enter the data into the computer we reviewed all the transcriptions of the interviews in which there were discrepancies in our coding. Now, with the project supervisor I came to realize how subjective the decision making can become. I found myself evaluating the interviews and trying to predict what the status would be according to prior decisions we both made and discussed. The frustration built because of having four "votes," we used the original two scores and then our current ones, and took the majority vote. Often the problem of

two for one status and two for another (or two for different or even four different) statuses arose.

I found myself truly puzzled about the significance of all studies involving clinical data and/or inter-rater reliability checks. In the many hours of listening and evaluating I feel I came to understand what the domains were and the typical responses I looked for in determining which status to choose. I would say that is the heart of the problem, it all comes down to the individual coder choosing. I see too many factors involved in the study I was involved in affecting the reliability. Until we can come up with a more consistent means of collecting data I don't see how the inter-rater reliability can be truly looked to as a part of the means for measuring subjective material in an objective manner. For this reason, Bennion and Adams (1986) have turned their attention to the development of a more objectifiable self-rating technique.

REFERENCES

Adams, G. R., Bennion, L., & Huh, K. (1987). Objective Measure of Ego Identity Status: A Reference Manual. Laboratory for Research on Adolescents, Utah State University, Logan, Utah 84322-2905.

Adams, G. R., Lee, Thomas, & Bennion, L. (1987). Training Manual for Interview Teams Utah Parent-Teen Relationship Project. Department of Family and Human Development, Utah State University, Logan, Utah, 84322-2905.

Adams, G. R. (1988). Personal communications.

Bennion, L. D., Adams, G. R. (1986). A Revision of the Extended Version of the Objective Measure of Ego Identity Status: An Identity Instrument for Use with Late Adolescents. Journal of Adolescent Research, 1, 183-198.

Figure 1.
Identity Status